Mahmoud, O., Harrison, A., Gul, A., Khan, Z., Metodiev, M. V., & Lausen, B. (2016). Minimizing redundancy among genes selected based on the overlapping analysis. In *Analysis of Large and Complex Data* (pp. 275-285). (Studies in Classification, Data Analysis, and Knowledge Organization). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-319-25226-1_24

Minimizing Redundancy among Genes Selected Based on the Overlapping Analysis

Osama Mahmoud[1,3], Andrew Harrison[1], Asma Gul[1], Zardad Khan[1], Metodi V. Metodiev[2], and Berthold Lausen[1]

[1] Department of Mathematical Sciences, University of Essex, United Kingdom.
ofamah@essex.ac.uk
[2] School of Biological Sciences/Proteomics Unit, University of Essex, UK.
[3] Department of Applied Statistics, Helwan University, Cairo, Egypt.

Abstract. For many functional genomic experiments, identifying the most characterizing genes is a main challenge. Both the prediction accuracy and interpretability of a classifier could be enhanced by performing the classification based only on a set of discriminative genes. Analyzing overlapping between gene expression of different classes is an effective criterion for identifying relevant genes. However, genes selected according to maximizing a relevance score could have rich redundancy. We propose a scheme for minimizing selection redundancy, in which the Proportional Overlapping Score (POS) technique is extended by using a recursive approach to assign a set of complementary discriminative genes. The proposed scheme exploits the gene masks defined by POS to identify more integrated genes in terms of their classification patterns. The approach is validated by comparing its classification performance with other feature selection methods, Wilcoxon Rank Sum, mRMR, MaskedPainter and POS, for several benchmark gene expression datasets using three different classifiers: Random Forest; k Nearest Neighbour; Support Vector Machine. The experimental results of classification error rates show that our proposal achieves a better performance.

Keywords

FEATURE SELECTION, CLASSIFICATION, PROPORTIONAL OVERLAPPING SCORES

1 Introduction

Microarray technology, as well as other high-throughput functional genomics experiments, have become a fundamental tool for gene expression analysis in recent years. A major challenge with microarray data is the problem of dimensionality; tens of thousands of genes' expressions are observed in a small number, tens to few hundreds, of observations. For a particular classification task, microarray data are inherently noisy since most genes are irrelevant and uninformative to the given classes (phenotypes).

Performing a supervised classification based on expressions of discriminative genes, identified by an effective gene selection technique, leads to

improved prediction accuracy, as well as interpretation of the biological relationship between genes and the considered clinical outcomes. This procedure of pre-selection of informative genes also helps in avoiding over-fitting problem and building a faster model by providing only the features that contribute most to the considered classification task. Identification of discriminative genes for their use in classification has been investigated in many studies (e.g., Apiletti et al. (2012), Mahmoud et al. (2014a)). Various approaches have been proposed including Best Individual Genes (Su et al. (2003)), Max-Relevance and Min-Redundancy based approaches (Peng et al. (2005)), Set Covering Machines (Kestler et al. (2006)), MaskedPainter (Apiletti et al. (2012)) and Proportional Overlapping Scores (POS) approach (Mahmoud et al. (2014a)). Different criteria have been used in order to detect the most informative genes including: p-values of statistical tests e.g. t-test or Wilcoxon rank sum test (Lausen et al. (2004)); ranking genes using statistical impurity measures e.g. information gain, gini index and max minority (Su et al. (2003)); selecting genes based on overlapping analysis (Apiletti et al. (2012), Mahmoud et al. (2014a)).

Analyzing the overlap between gene expression measures for different classes is an effective criterion for identifying discriminative genes for a considered classification task. Mahmoud et al. (2014a) developed a procedure specifically designed to select genes based on their overlapping degree across different classes. This procedure, named *Proportional Overlapping Score* (POS), calculates a relevance score for each gene. For binary class situations, this score estimates the overlapping degree between the expressions intervals of both classes taking into account three factors that form the characteristics of classes' overlapping. It has been defined to provide higher scores for genes with lower discriminative power. Genes are then ranked in ascending order according to their scores. POS method characterizes each gene by means of a *gene mask* that represents the capability of a gene to unambiguously assign training observations to their correct classes. Characterization of genes using training observation masks with their overlapping scores allow the detection of a minimum subset of genes that provides the best classification coverage on a training set of observations. A final gene set is then provided by combining the minimum gene subset with the top ranked genes according to the estimated scores. Feature selection produced by POS is robust against outliers, since gene masks are defined based on the interquartile range of gene's expressions. However, the top ranked genes, given based on POS relevance score, may provide a classifier with redundant information.

In this article, we propose an extended version of POS method, called POSr, that can exploit detection of the minimum subset of genes in a recursive way in order to mitigate redundancy in the final gene selection.

The article is organized as follows. Section 2 shows the main idea of POS and explains the proposed method. The results of proposal are compared with some other gene selection techniques in section 3. Section 4 concludes the article.

2 Methods

2.1 Overlapping Analysis for Binary Class Problems

Microarray data are usually presented in the form of a gene expression matrix, $X = \left[x_{ij}\right]$, such that $X \in \mathbb{R}^{P \times N}$ and x_{ij} is the observed expression value of gene i for observation (tissue sample) j where $i = 1, \ldots, P$ and $j = 1, \ldots, N$. Each observation is also characterized by a target class label, y_j, representing the phenotype of the observation being studied. Let $Y \in \mathbb{R}^N$ be the vector of class labels such that its jth element, y_j, has a single value c which is either 1 or 2.

Analyzing the overlap between expression intervals of a gene for different classes can provide a classifier with an important aspect of a gene's characteristic. The idea is that a certain gene i can assign observations to class c because their gene i expression interval in that class is not overlapping with gene i interval of the other class. In other words, gene i has the ability to correctly classify observations for which their gene i expressions fall within the expression interval of a single class.

POS method, proposed by Mahmoud et al. (2014a), initially exploits the interquartile range approach to robustly define gene masks that report the discriminative power of genes avoiding outlier effects. Construction of these masks can be described as follows.

Core Intervals and Gene Masks
For a certain gene i, two expression intervals, one for each class, can be defined for that gene. The cth class core interval for gene i can be defined in the form:

$$I_{i,c} = [a_{i,c}, b_{i,c}], \quad i = 1, \ldots, P, \quad c = 1, 2, \tag{1}$$

such that:

$$a_{i,c} = Q_1^{(i,c)} - 1.5\, IQR^{(i,c)}, b_{i,c} = Q_3^{(i,c)} + 1.5\, IQR^{(i,c)}, \tag{2}$$

where $Q_1^{(i,c)}$, $Q_3^{(i,c)}$ and $IQR^{(i,c)}$ denote the first, third empirical quartiles, and the interquartile range of gene i expression values for class c respectively. The multiplier value of 1.5 is the default value that commonly used with the interquartile range approach for detecting outliers (Tukey 1977).

For each gene, a mask is defined based on its observed expression values and constructed core intervals. Gene i mask is represented by a vector of length equal to the total number of observations. It reports the observations that gene i can unambiguously assign to their correct target classes. Thus, gene masks can represent the capability of genes to classify correctly each observation, i.e. it represents a gene's classification power. For a particular gene i, element j of its mask is set to 1 if the corresponding expression value x_{ij} belongs only to core expression interval I_{i,c_j} of the single class c_j, where c_j is the target class of observation j. Otherwise, it is set to zero.

Figure 1 shows the constructed core expression intervals $I_{i,1}$ and $I_{i,2}$ associated with a particular gene i along-with its gene mask. The non-overlapped

observations are represented by circles. The gene mask is sorted corresponding to the observations ordered by increasing expression values.

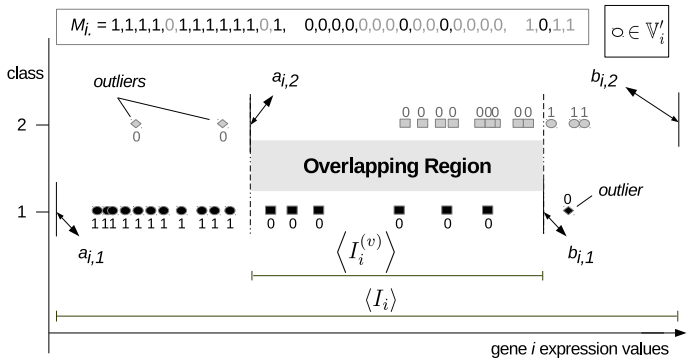

Fig. 1. Core intervals with gene mask. An example for core expression intervals of a gene with 18 and 14 observations belonging to class 1, in black colour, and class 2, in grey colour, respectively, with its associated mask elements. Elements of the non-overlapped observations set are represented by circles.

A matrix of gene masks $M = \left[m_{ij}\right]$ can be produced such that the mask of gene i is presented by $M_{i.}$(the ith row of M) and gene mask element m_{ij} is defined as:

$$m_{ij} = \begin{cases} 1 & if \quad j \in \mathbb{V}'_i \\ 0 & otherwise \end{cases}, \tag{3}$$

where $\mathbb{V}'_i$ is the set that includes non-outliers observations whose observed expressions fall into the non-overlapping region such that $i = 1, \ldots, P$ and $j = 1, \ldots, N$.

Proportional Overlapping Score

An overlapping measure, called proportional overlapping score (*POS*), is developed to estimate the overlapping degree between different expression intervals taking into account three factors: (1) length of the overlapping region; (2) number of overlapped observations; (3) the proportion of classes' contribution to the overlapped observations (Mahmoud et al. (2014a)). For each gene i, POS_i is estimated as follows:

$$POS_i = 4 \, \frac{\left\langle I_i^{(v)} \right\rangle}{\langle I_i \rangle} \, \frac{\nu_i}{\ell_i} \left(\prod_{c=1}^{2} \theta_c \right), \tag{4}$$

where $\left\langle I_i^{(v)} \right\rangle$ is the length of the overlapping region, and $\langle I_i \rangle$ is the length of the total core interval which is given by the region between the global minimum and global maximum boundaries of core intervals for both classes, see Figure 1. Whereas ν_i and ℓ_i represent number of observations whose observed expressions of gene i fall within the overlapping region and number

of non-outlier observations respectively, while θ_c is the proportion of class c observations among overlapped observations. Hence, θ_c can be defined as:

$$\theta_c = \frac{\nu_{i,c}}{\nu_i}, \tag{5}$$

where $\nu_{i,c}$ represents number of the overlapped observations belonging to class c. The factor 4 is included in (4) in order to scale *POS* values within the interval [0, 1] (Mahmoud et al. (2014a)). According to (4) and (5), the value of *POS* measure for gene i shown in Figure 1 is $4 \cdot \frac{15}{29} \cdot \left(\frac{6}{15} \cdot \frac{9}{15}\right) \cdot \frac{\left\langle I_i^{(v)} \right\rangle}{\langle I_i \rangle} = \frac{72}{145} \cdot \frac{\left\langle I_i^{(v)} \right\rangle}{\langle I_i \rangle}$.

2.2 Recursive Minimum Sets for Minimizing Redundancy (POSr)

POS gives its final selection by combining a minimum gene subset produced using gene masks, defined in (3), with the top ranked genes according to the estimated *POS* scores, defined in (4). It is an effective feature selection method for identifying discriminative genes for a considered classification task.

However, POS selections may provide a classifier with redundant information since the set of top ranked genes is likely to have redundancy among its members. Such a redundancy increases the model complexity since it increases the dimensionality without adding further information. Moreover, redundancy may affect classification prediction accuracy as well as interpretation of the underlying biological relationship between the features and considered clinical outcomes.

A gene mask reflects the capability of the gene to correctly classify each observation to its target class. Genes with higher number of 1 bits in their masks are more informative to the considered classification problem (see (3)). When two genes classify in the same way the same observations, then their masks should be identical. Genes with complementary masks, on the other hand, can provide diverse information to the classifier model.

In this article, we propose an extended version of POS, called POSr, in which gene masks along-with *POS* measure are exploited to identify minimum subsets of genes in a recursive way in order to mitigate the potential redundancy in the final gene selection. The subset is designated to be the minimum one that correctly classify the maximum number of observations in a given training set, avoiding the effects of expression outliers.

Let $\mathbb{G}_z$ be a set of remaining genes at the zth iteration given by excluding the selected subset of genes at the $(z-1)$th iteration, such that $\mathbb{G}_1$ is the full set of all genes (i.e., $|\mathbb{G}_1| = P$). Also, let $\overline{M}(\mathbb{G}_z)$ be its aggregate mask which is defined as the logical disjunction *(logic OR)* among all masks corresponding to genes that belong to the set $\mathbb{G}_z$. It can be expressed as follows:

$$\overline{M}(\mathbb{G}_z) = \bigvee_{i \in \mathbb{G}_z} M_{i.} \tag{6}$$

At iteration z, our objective is to search the set, $\mathbb{G}_z$, for the minimum subset, denoted by $\mathbb{G}_z^*$, for which $\overline{M}(\mathbb{G}_z^*)$ equals to the aggregate mask of the

corresponding set of genes, $\overline{M}(\mathbb{G}_z)$. In other words, our minimum subset of genes should satisfy the following statement:

$$\underset{\mathbb{G}_z^* \subseteq \mathbb{G}_z}{argmin} \left(\left| \mathbb{G}_z^* \right| \left| \left(\overline{M}(\mathbb{G}_z^*) = \bigvee_{i \in \mathbb{G}_z^*} M_{i.} = \overline{M}(\mathbb{G}_z) \right) \right. \right). \tag{7}$$

This procedure is performed in a recursive way and ends when the required number of genes, set by the user, are selected.

The pseudo code of our procedure, POSr, is reported in Algorithm 1. Its inputs are: the matrix of gene masks, M; POS scores for all genes; number of genes to be selected, r. It produces the sequence of selected genes, $\mathbb{T}^*$, as output.

Algorithm 1 POSr Method: Recursive Minimum Subsets

Inputs: M, POS scores and number of required genes (r).
Output: Sequence of the selected genes $\mathbb{T}^*$.

1: $z = 0$ {Initialization}
2: $\mathbb{T} = \emptyset$
3: **while** $\left|\mathbb{T}\right| < r$ **do**
4: $z = z + 1$
5: $k = 0$ {Initialization of individual selection}
6: $\mathbb{G}_z^* = \emptyset$
7: $\overline{M}(\mathbb{G}_z^*) = \mathbf{0}_N$
8: **while** $\overline{M}(\mathbb{G}_z^*) \neq \overline{M}(\mathbb{G}_z)$ **do**
9: $k = k + 1$
10: $\mathbb{S}_{zk} = \underset{i \in \mathbb{G}_z}{argmax} \left(\sum_{j=1}^{N} I\left(m_{ij}^{(k)} = 1\right) \right)$ {Assign gene set whose masks have max. bits of 1}
11: $g_{zk} = \underset{i \in \mathbb{S}_{zk}}{argmin}(POS_i)$ {Select the candidate with the best score among the assigned set}
12: $\mathbb{G}_z^* = \mathbb{G}_z^* + g_{zk}$ {Update the target set by adding the selected candidate}
13: **for all** $i \in \mathbb{G}_z$ **do**
14: $M_{i.}^{(k+1)} = M_{i.}^{(k)} \wedge \overline{M}'(\mathbb{G}_z^*)$ {update gene masks such that the uncovered observations are only considered}
15: **end for**
16: **end while**
17: $\mathbb{T} = \mathbb{T} + \mathbb{G}_z^*$
18: $\mathbb{G}_{z+1} = \mathbb{G}_z - \mathbb{G}_z^*$
19: **end while**
20: $\mathbb{T}^*$ is the sequence whose members are the first r genes in $\mathbb{T}$
21: **return** $\mathbb{T}^*$

At the initial step ($z = 0$), we let $\mathbb{T} = \emptyset$ (line 2); where $\mathbb{T}$ is a set created to contain the successively selected minimum subsets of genes. Then at each iteration, z, the following steps are performed:

1. We let $k = 0$, $\mathbb{G}_z^* = \emptyset$ and $\overline{M}(\mathbb{G}_z^*) = \mathbf{0}_N$ (lines 5-7) to initialize individual selection within the minimum subset $\mathbb{G}_z^*$, where $\overline{M}(\mathbb{G}_z^*)$ is the aggregate mask of the set $\mathbb{G}_z^*$, see (6). Then at each sub-iteration, k, the following sub-steps are performed:

a) Among genes of the set $\mathbb{G}_z$, the one(s) with the highest number of mask bits assigned to 1 is (are) chosen to form the set $\mathbb{S}_{zk}$ (line 10).
b) The gene with the lowest *POS* score among genes in $\mathbb{S}_{zk}$, if there are more than one, is then selected (line 11). It is denoted by g_{zk}.
c) The set $\mathbb{G}_z^*$ is updated by adding the selected gene, g_{zk} (line 12).
d) All masks of genes in $\mathbb{G}_z$ are also updated by performing the logical conjunction (*logic AND*) with negated aggregate mask of set $\mathbb{G}_z^*$ (line 14). Note that $M_{i.}^{(k)}$ represents updated mask of gene i at the kth iteration such that $M_{i.}^{(1)}$ is its original gene mask whose elements are computed according to (3).
e) This sub-procedure is successively iterated and ends when all masks of genes in $\mathbb{G}_z$ have no one bits anymore, i.e. the selected genes cover the maximum number of observations. This situation is accomplished iff $\overline{M}(\mathbb{G}_z^*) = \overline{M}(\mathbb{G}_z)$.

2. The set $\mathbb{T}$ is updated by adding the detected minimum subset of genes, $\mathbb{G}_z^*$ (line 17).
3. Genes within the selected minimum subset, $\mathbb{G}_z^*$, are then removed from the set of genes, $\mathbb{G}_z$ (line 18).
4. The procedure is successively iterated and ends when the size of the set $\mathbb{T}$ is greater than or equal the number of required genes, r. Then, the target sequence of selected genes, $\mathbb{T}^*$, is produced by selecting the first r genes in $\mathbb{T}$ (lines 20, 21).

Thus, this approach combines recursively the detected minimum subsets of genes that provide the best classification coverage for a given training set. Selection of the minimum subsets based on the updated gene masks allows to minimize redundancy among the final selection list.

3 Results and Discussion

For evaluating a feature selection method, one can assess the accuracy of a classifier applied after the feature selection process. Such an assessment can verify the efficiency of gene selections. In this article, our experiment is conducted using seven publicly available gene expression datasets in which the POSr method is validated by comparison with three well-known gene selection techniques along-with POS method. The performance is evaluated by obtaining the classification error rates from three different classifiers: Random Forest (RF); k Nearest Neighbor (kNN); Support Vector Machine (SVM).

Table 1 summarizes the characteristics of the datasets. The estimated classification error rate is based on the Random Forest classifier with the full set of features, without pre-selection.

Fifty repetitions of 10-fold cross validation analysis were performed for each combination of dataset, feature selection algorithm, and a given number of selected genes, up to 50, with the considered classifiers. For each experimental repetition, the split seed was changed while the same folds and

Table 1. Description of used gene expression datasets.

Dataset	*Genes*	*Observations*	*Class-sizes*	*Est. Error*	*Source*
Leukaemia	7129	72	47/25	0.049	Golub et al. (1999)
Breast	4948	78	34/44	0.369	Michiels et al. (2005)
Srbct	2308	54	29/25	0.0008	Statnikov et al. (2005)
Lung	12533	181	150/31	0.003	Gordon et al. (2002)
GSE24514	22215	49	34/15	0.0406	Alhopuro et al. (2012)
GSE4045	22215	37	29/8	0.2045	Laiho et al. (2007)
GSE14333	54675	229	138/91	0.4141	Jorissen et al. (2009)

training datasets were kept for all feature selection methods. To avoid bias, gene selection algorithms have been performed only on the training sets. For each fold, the best subset of genes has been selected according to the Wilcoxon Rank Sum technique (Wil-RS), Minimum Redundancy Maximum Relevance (mRMR) method (Peng et al. (2005)), MaskedPainter (MP) (Apiletti et al. (2012)), Proportional Overlapping Scores (POS) (which is implemented in *propOverlap* R package (Mahmoud et al. (2014b)), along-with our proposed method. The expressions of the selected genes as well as the class labels of the training observations have then been used to construct the considered classifiers. The classification error rates on the test sets are separately reported for each classifier and the average error rate over all the fifty repetitions is then computed.

To highlight the entire performances of the compared methods against our proposed approach, a comparison between the minimum error rates achieved by each method was conducted. Table 2 summarizes these results. Each row shows the minimum error rate (along-with its corresponding set size, shown in brackets) for a specific dataset, reported in the first column. In addition, the error rates of the corresponding classifiers with the full set of features, without feature selection, are reported in the last column. Due to limitations of the R package 'mRMRe' (De Jay et al. (2013)), mRMR selections could not be conducted for datasets having more than '46340' features. Therefore, mRMR method is excluded from the analysis of the 'GSE14333' dataset.

Table 2 demonstrates that the proposed approach, POSr, provides the minimum error rates (the highest accuracy) for all used classifier models with most of the used datasets. In particular, for the 'Leukaemia', 'Lung' and 'GSE4045' datasets, it outperforms the other methods using all different classifiers. For the 'Breast' and 'Srbct' datasets, POSr provide the best performance using kNN and SVM, for the 'Breast' dataset, and RF, for the 'Srbct' dataset. While, on the 'GSE14333' and 'GSE24514' datasets, WilRS and POS methods respectively outperformed the other compared methods.

Table 2. Comparison between the minimum error rates yielded by the feature selection methods using RF, *k*NN and SVM classifiers.

Dataset	*Classifier*	*Wil-RS*	*mRMR*	*MP*	*POS*	*POSr*	*Full Set*
	RF	0.030 (20)	0.118 (40)	0.015 (9)	0.0002 (40)	**0.000 (9)**	0.049
Leukaemia	*k*NN	0.074 (6)	0.135 (50)	0.019 (1)	**0.005 (1)**	**0.005 (1)**	0.109
	SVM	0.047 (8)	0.126 (50)	0.022 (1)	**0.005 (1)**	**0.005 (1)**	0.131
	RF	0.040 (30)	0.016 (48)	0.008 (46)	0.007 (48)	**0.006 (48)**	0.003
Lung	*k*NN	0.203 (12)	0.027 (49)	0.017 (17)	0.011 (12)	**0.002 (40)**	0.0005
	SVM	0.066 (50)	0.026 (50)	0.021 (19)	0.010 (47)	**0.008 (38)**	0.024
	RF	0.371 (50)	0.407 (48)	0.354 (48)	**0.308 (45)**	0.317 (48)	0.369
Breast	*k*NN	0.405 (11)	0.404 (50)	0.346 (19)	0.332 (11)	**0.328 (11)**	0.405
	SVM	0.401 (39)	0.407 (50)	0.359 (21)	0.313 (22)	**0.303 (37)**	0.438
	RF	0.069 (24)	0.074 (46)	0.009 (32)	0.003 (48)	**0.002 (44)**	0.0008
Srbct	*k*NN	0.157 (3)	0.098 (48)	0.005 (26)	**0.005 (22)**	0.008 (32)	0.034
	SVM	0.131 (50)	0.124 (49)	0.010 (21)	**0.003 (8)**	0.004 (47)	0.079
	RF	0.134 (24)	0.187 (37)	0.137 (21)	0.114 (27)	**0.105 (33)**	0.205
GSE4045	*k*NN	0.166 (43)	0.207 (38)	0.137 (50)	0.142 (3)	**0.112 (6)**	0.103
	SVM	0.134 (24)	0.187 (37)	0.095 (47)	0.114 (29)	**0.085 (47)**	0.214
	RF	**0.421 (10)**	-	0.438 (31)	0.437 (34)	0.442 (44)	0.414
GSE14333	*k*NN	**0.420 (8)**	-	0.455 (23)	0.450 (34)	0.448 (47)	0.438
	SVM	0.427 (9)	-	**0.412 (1)**	0.431 (1)	0.431 (1)	0.407
	RF	0.054 (47)	0.063 (50)	0.036 (48)	**0.032 (24)**	0.034 (26)	0.041
GSE24514	*k*NN	**0.032 (20)**	0.041 (50)	0.036 (50)	0.039 (50)	0.038 (49)	0.041
	SVM	0.041 (40)	0.059 (50)	0.037 (40)	**0.034 (30)**	0.036 (43)	0.070

Boldface numbers indicate the lowest classification error rates (highest accuracy among compared methods) achieved using the corresponding classifier. The numbers in brackets represent the size of the gene sets that corresponding to the minimum error rate.

Figure 2 shows that our proposed approach provides less classification error rates than other compared gene selection methods on the 'Breast' and 'Lung' datasets at different selected gene set sizes. The stability index proposed by Lausser et al. (2013) is used to measure the stability of the compared method at different set sizes of features. The relation between the accuracy and stability has been depicted for the 'Lung' dataset. Different dots for the same gene selection method correspond to different set sizes of genes. For all classifiers, POSr achieves a good trade-off between accuracy and stability for 'Lung' data, see the second row panels of Figure 2.

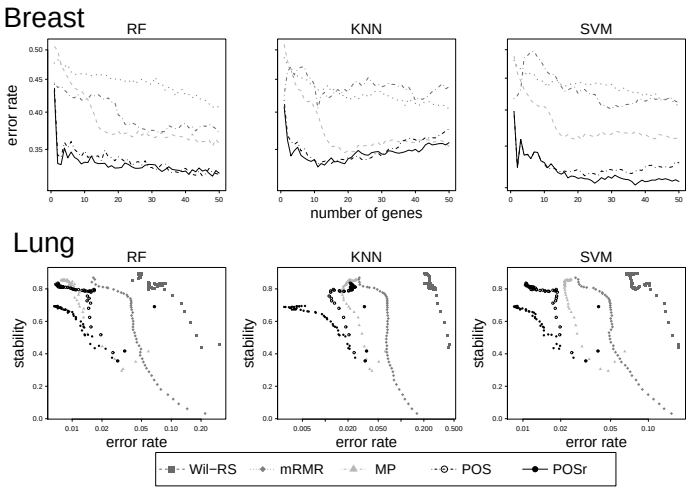

Fig. 2. Averages of classification error rates and stability-accuracy plots: (*first row*) averages of classification error rates for the 'Breast' dataset using RF, kNN and SVM classifiers, (*second row*) stability-accuracy plots for the 'Lung' dataset.

4 Conclusion

A gene selection method, POSr, is proposed as an extension of Proportional Overlapping Scores (POS) technique. The proposed approach detects minimum subsets of genes in a successive way. The final selection is then produced by combining these subsets in order to reduce the redundancy among selected genes. It is designed for binary class situations. The classification error rates achieved by Random Forest, *k* Nearest Neighbour and Support Vector Machine classifiers for POSr were compared with Wilcoxon Rank Sum, Maximum Relevance Minimum Redundancy, MaskedPainter and POS on seven benchmarked gene expression datasets. The relation between classification accuracy and selection stability is also outlined. The proposed method performed better than compared methods on most data sets for all classifiers. It is an effective approach in enhancing the prediction classification performance of the considered classifier models using less number of features compared to the other studied gene selection methods. Furthermore, POSr approach provides good stability scores at small as well as large sets of selected genes.

References

ALHOPURO, P., SAMMALKORPI, H., NIITTYMÄKI, I., BISTRÖM, M., RAITILA, A. et al. (2012): Candidate Driver Genes In Microsatellite-Unstable Colorectal Cancer. *Int. J. Cancer, 130(7): 1558–1566*.

APILETTI, D., BARALIS, E., BRUNO, G. and FIORI, A. (2012): Maskedpainter: Feature Selection For Microarray Data Analysis. *Intell Data Anal, 16(4):717–737*.

DE JAY, N., PAPILLON-CAVANAGH, S., OLSEN, C., EL-HACHEM, N., BONTEMPI, G. and HAIBE-KAINS, B. (2013): mrmre: An R Package For Parallelized Mrmr Ensemble Feature Selection. *Bioinformatics, 29(18): 2365–2368*.

GOLUB, TR., SLONIM, DK., TAMAYO, P., HUARD, C., GAASENBEEK, M. et al. (1999): Molecular Classification Of Cancer: Class Discovery And Class Prediction By Gene Expression Monitoring. *Science, 286(5439):531–537*.

GORDON, G., JENSEN, R., HSIAO, L., GULLANS, S., BLUMENSTOCK, E. et al. (2002): Translation Of Microarray Data Into Clinically Relevant Cancer Diagnostic Tests Using Gene Expression Ratios In Lung Cancer And Mesothelioma. *Cancer Resh., 62(17): 4963–4967*.

JORISSEN, RN., GIBBS, P., CHRISTIE, M., PRAKASH, S., LIPTON, L. et al. (2009): Metastasis-Associated Gene Expression Changes Predict Poor Outcomes In Patients With Dukes Stage B And C Colorectal Cancer. *Clinical Cancer Res., 15(24): 7642–7651*.

KESTLER, H., LINDNER, W. and MüLLER, A. (2006): Learning and feature selection using the set covering machine with data-dependent rays on gene expression profiles. In SCHWENKER, F. and MARINAI, S., editors, *Artificial Neural Networks in Pattern Recognition (ANNPR 06) volume LNAI 4087, pages 286297.*, Springer-Verlag, Heidelberg.

LAIHO, P., KOKKO, A., VANHARANTA, S., SALOVAARA, R., SAMMALKORPI, H. et al. (2007): Serrated Carcinomas Form A Subclass Of Colorectal Cancer With Distinct Molecular Basis. *Oncogene, 26(2): 312–320*.

LAUSEN, B., HOTHORN, T., BRETZ, F. and SCHUMACHER, M. (2004): Assessment Of Optimal Selected Prognostic Factors. *Biom J, 46(3):364–374*.

LAUSSER, L., MÜSSEL, C., MAUCHER, M. and KESTLER, HA. (2013): Measuring And Visualizing The Stability Of Biomarker Selection Techniques. *Comput. Stat., 28(1): 51–65*.

MAHMOUD, O., HARRISON, A., PERPEROGLOU, A., GUL, A., KHAN, Z., METODIEV, M. and LAUSEN, B. (2014a): A Feature Selection Method For Classification Within Functional Genomics Experiments Based On The Proportional Overlapping Score. *BMC Bioinformatics, 15:274*.

MAHMOUD, O., HARRISON, A., PERPEROGLOU, A., GUL, A., KHAN, Z. and LAUSEN, B. (2014b): propOverlap: Feature (gene) selection based on the Proportional Overlapping Scores. R package version 1.0, http://CRAN.R-project.org/package=propOverlap.

MICHIELS, S., KOSCIELNY, S. and HILL, C. (2005): Prediction Of Cancer Outcome With Microarrays: A Multiple Random Validation Strategy. *The Lancet, 365(9458): 488–492*.

PENG, H., LONG, F. and DING, C. (2005): Feature Selection Based On Mutual Information Criteria Of Max-Dependency, Max-Relevance, And Min-Redundancy. *Pattern Anal Mach Intell IEEE Trans, 27(8):1226–1238*.

STATNIKOV, A., ALIFERIS, CF., TSAMARDINOS, I., HARDIN, D. and LEVY, S. (2005): A Comprehensive Evaluation Of Multicategory Classification Methods For Microarray Gene Expression Cancer Diagnosis. *Bioinformatics, 21(5): 631–643*.

SU, Y., MURALI, T., PAVLOVIC, V., SCHAFFER, M. and KASIF, S. (2003): Rankgene: Identification Of Diagnostic Genes Based On Expression Data. *Bioinformatics, 19(12):1578–1579*.

TUKEY, J. (1977): Exploratory data analysis. *Reading, Mass*.